

Constraining duty cycles through a Bayesian technique

P. Romano¹, C. Guidorzi², A. Segreto¹, L. Ducci^{3,4}, and S. Vercellone¹

¹ INAF, Istituto di Astrofisica Spaziale e Fisica Cosmica - Palermo, Via U. La Malfa 153, I-90146 Palermo, Italy
e-mail: romano@ifc.inaf.it

² Dipartimento di Fisica e Scienze della Terra, Università di Ferrara, Via Saragat 1, I-44122 Ferrara, Italy

³ Institut für Astronomie und Astrophysik, Eberhard Karls Universität, Sand 1, 72076 Tübingen, Germany

⁴ ISDC Data Center for Astrophysics, Université de Genève, 16 chemin d'Écogia, 1290 Versoix, Switzerland

Received 11 May 2014; accepted 2014 October 15

ABSTRACT

The duty cycle (DC) of astrophysical sources is generally defined as the fraction of time during which the sources are active. It is used to both characterize their central engine and to plan further observing campaigns to study them. However, DCs are generally not provided with statistical uncertainties, since the standard approach is to perform Monte Carlo bootstrap simulations to evaluate them, which can be quite time consuming for a large sample of sources. As an alternative, considerably less time-consuming approach, we derived the theoretical expectation value for the DC and its error for sources whose state is one of two possible, mutually exclusive states, inactive (off) or flaring (on), as based on a finite set of independent observational data points. Following a Bayesian approach, we derived the analytical expression for the posterior, the conjugated distribution adopted as prior, and the expectation value and variance. We applied our method to the specific case of the inactivity duty cycle (IDC) for supergiant fast X-ray transients, a subclass of flaring high mass X-ray binaries characterized by large dynamical ranges. We also studied IDC as a function of the number of observations in the sample. Finally, we compare the results with the theoretical expectations. We found excellent agreement with our findings based on the standard bootstrap method. Our Bayesian treatment can be applied to all sets of independent observations of two-state sources, such as active galactic nuclei, X-ray binaries, etc. In addition to being far less time consuming than bootstrap methods, the additional strength of this approach becomes obvious when considering a well-populated class of sources ($N_{\text{src}} \geq 50$) for which the prior can be fully characterized by fitting the distribution of the observed DCs for all sources in the class, so that, through the prior, one can further constrain the DC of a new source by exploiting the information acquired on the DC distribution derived from the other sources.

Key words. Methods: statistical – Methods: numerical – Methods: observational – X-rays: binaries

1. Introduction

In astrophysics it is often crucial to determine the duty cycle (DC) of a source, or a class of sources, in order to understand both their central engines and to plan additional observing campaigns aiming at best studying them. Generally, the DC is defined as the fraction of time, usually expressed in percentages, during which the source is active, or

$$\text{DC} = T_{\text{active}}/T_{\text{Tot}}, \quad (1)$$

where T_{active} is the time spent above some instrumental threshold or some scientifically interesting flux value, and T_{Tot} is the total exposure. In the case of periodic sources, such as classical X-ray binaries, T_{active} is generally the time during which an n - σ detection is achieved (n being 3 or 5, depending on the detection method), and T_{Tot} is the orbital period P_{orb} or the spin period P_{spin} (e.g. Henry & Paik 1969; Fragos et al. 2009; Kneivitt et al. 2014).

For active galactic nuclei (AGNs), the DC is often defined as the fraction of time a source spends in a flaring state, that is, at n times the average flux, \bar{F} , with n being a small number, depending on the purpose of the study (e.g. Jorstad et al. 2001; Vercellone et al. 2004; Ackermann et al. 2011). For example, in Vercellone et al. (2004) the DC is defined as $\chi = \frac{\tau}{\tau+T}$, where T is the time spent in a low flux level (*off state*) and τ is the time

spent in a high flux level (*on state*), defined by $HSN = \sum_{i=1}^n C_i$, where $C_i = 1$ if $F_i \geq 1.5 \times \bar{F}$ and $C_i = 0$ otherwise.

Alternatively, when a source shows a very large dynamical range (a few orders of magnitude), more can be inferred about its nature by considering the *inactivity duty cycle* (IDC, Romano et al. 2009) defined as the time a source remains undetected down to a certain flux limit F_{lim} ,

$$\text{IDC} = \Delta T_{\Sigma}/[\Delta T_{\text{tot}}(1 - P_{\text{short}})], \quad (2)$$

where ΔT_{Σ} is the sum of the exposures accumulated in all observations where only a 3σ upper limit was achieved, ΔT_{tot} is the total exposure accumulated, and P_{short} is the percentage of time lost to short observations that need to be discarded in order to differentiate between non-detections due to lack of exposure from non-detections due to a true low flux state.

Since DCs (and IDCs) are integral quantities depending on the total observing time and the total time spent above (or below) a given flux threshold, they are implicitly dependent on the instrumental sensitivity, observing coverage, and the characteristic source variability timescales. The implicit assumption is that, in order to obtain a meaningful DC, the observations used to calculate them are independent, that is, each observation is not triggered by the previous ones. This is the case, for example, of monitoring programmes whose monitoring pace and exposures are defined a priori and do not depend on the source state.

Table 1. Source sample properties and comparison of measured IDCs with Bayesian estimates and MC simulations.

Source	Orbital period (d)	Observation N	Type	IDC ^a (%)	Bayesian method confidence intervals ^b			Monte Carlo simulations	
					68.3 %	95.4 %	99.7 %	$IDC_{\text{sim}} \pm s_{\text{sim}}^c$ (%)	S_a^d
IGR J08408–4503	–	77	Y	67.2	61.5–72.1	55.8–76.8	50.1–81.2	67.3 ± 5.6	40
IGR J16328–4726	10.076	94	Y	61.0	55.8–65.8	50.7–70.4	45.5–74.8	61.0 ± 5.6	40
IGR J16465–4507	30.243	61	Y	5.1	3.5–9.5	1.8–13.9	0.8–19.2	5.2 ± 2.9	40
IGR J16479–4514	3.3193	139	Y	19.4	16.5–23.2	13.6–26.9	11.0–30.9	19.4 ± 3.6	80
XTE J1739–302	51.47	181	Y	38.8	35.3–42.5	31.9–46.2	28.5–49.9	39.0 ± 4.7	70
IGR J17544–2619	4.926	138	Y	54.5	50.2–58.6	46.0–62.7	41.8–66.7	54.5 ± 5.3	50
AX J1841.0–0536	–	87	Y	28.4	24.1–33.7	19.8–38.9	16.0–44.2	28.5 ± 5.6	40
IGR J16418–4532	3.73886	15	O	11.0	7.2–24.1	3.0–36.1	0.9–49.1	11.3 ± 8.0	–
IGR J17354–3255	8.448	22	O	33.4	25.1–44.5	17.2–54.9	11.0–64.8	33.3 ± 10.4	–
IGR J18483–0311	18.545	23	O	26.6	19.5–37.5	12.7–47.6	7.6–57.8	26.7 ± 9.4	–

Notes. ^(a) From Eq. 2 (see Romano et al. 2014, and references therein). ^(b) Theoretical confidence intervals of IDC (Sect. 2, Eq. 7). ^(c) Simulated sample mean and standard variance (Sect. 3, $M = 10^4$ data sets drawn from the observed sample of size N). ^(d) Minimum number of observations required for an IDC with the desired accuracy (Sect. 3.1, $M = 10^4$ data sets, drawn from a sample of size $S = 10, 20, 30, \dots, N$).

In this paper we determine the theoretical expectation value of DC and its error. We then consider one specific case, the IDCs measured from ten *Swift* (Gehrels et al. 2004) X-ray Telescope (XRT, Burrows et al. 2005) observing campaigns on supergiant fast X-ray transients (SFXTs), a subclass of high mass X-ray binaries known for their rapid hard X-ray flaring behaviour and large dynamical range (up to 5 orders of magnitude), and compare the theoretical expectations with both the observed values and with those obtained from Monte Carlo simulations. We also evaluate how the IDC varies as a function of the number of observations available and estimate how many observations are required to obtain an IDC within a desired accuracy. Finally, we supply the reader with useful R-language (R Core Team 2014), IDL, and C-language procedures to calculate several confidence intervals (c.i.) on the DC estimate for a given source.

2. Statistical estimate of the duty cycle

We consider one source for which N independent observations were collected and for which the DC was calculated as described in Sect. 1. In the following we estimate the DC, that we hereafter define as μ , but the formalism is unchanged for the case of the IDC which we consider in Sect. 3. In all generality, the stochastic variable *state of the source* can be seen as a discrete random variable that can take only one of two possible, mutually exclusive states, active (off) and flaring (on), so that μ is the probability of finding the source active in a given casual pointing. After N observations, the probability of finding the source active m times is given by the binomial

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}, \quad (3)$$

with an expectation value $E\{m\} = \mu N$ and variance $\text{var}\{m\} = N\mu(1 - \mu)$. Once N and m are known, where $m = N\mu_{\text{est}}$ and μ_{est} is the DC measured from the N observations, then the problem becomes estimating the statistical offset of μ from μ_{est} . From the central limit theorem μ_{est} is normally distributed in the limit of large values of m and $N - m$ with $E\{\mu_{\text{est}}\} = \mu$ and $\sigma\{\mu_{\text{est}}\} = \sqrt{\mu(1 - \mu)/N}$.

Hereafter, we adopt a Bayesian treatment, in which μ is treated as a random variable whose probability density function

(PDF) depends on the observed values for N and m . From the Bayes theorem, the posterior distribution is proportional to the product of the likelihood and the prior function, and the posterior distribution is to all intents and purposes a PDF of the random variable μ given the observed values for N and m ,

$$P(\mu|N, m) \propto P(m|N, \mu) \cdot p(\mu), \quad (4)$$

where $P(m|N, \mu)$ is the likelihood given by Eq. (3) and is meant to be a function of μ . The prior is denoted by $p(\mu)$. Apart from a normalization term, the likelihood is the Beta distribution of μ given N and m

$$\text{Beta}(\mu|N, m) = \frac{\Gamma(N+2)}{\Gamma(m+1)\Gamma(N-m+1)} \mu^m (1 - \mu)^{N-m}. \quad (5)$$

The convenient choice (Bishop 2006, Sect. 2.2.1) for a prior function is a conjugated distribution, which is the Beta distribution with parameters a and b ,

$$p(\mu|a, b) = \text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1 - \mu)^{b-1}. \quad (6)$$

After proper normalization, the posterior in Eq. (4) becomes

$$P(\mu|m, N, a, b) = \frac{\Gamma(N+a+b)}{\Gamma(m+a)\Gamma(N-m+b)} \mu^{m+a-1} (1 - \mu)^{N-m+b-1}. \quad (7)$$

From Eq. (7) the expectation value and variance of μ are

$$E\{\mu\} = \frac{m+a}{N+a+b} = \frac{\mu_{\text{est}} + a/N}{1 + (a+b)/N} \quad (8)$$

$$\begin{aligned} \text{var}\{\mu\} &= \frac{(m+a)(N-m+b)}{(N+a+b)^2(N+a+b+1)} \\ &= \frac{(\mu_{\text{est}} + a/N)(1 - \mu_{\text{est}} + b/N)}{(1 + (a+b)/N)^2(N+a+b+1)}. \end{aligned} \quad (9)$$

The case of an uninformative prior is easily recovered for $a = b = 1$. We note that in the asymptotic limit of large values of N ,

$$E\{\mu\} \simeq \mu_{\text{est}} \quad (10)$$

$$\text{var}\{\mu\} \simeq \frac{\mu_{\text{est}}(1 - \mu_{\text{est}})}{N}, \quad (11)$$

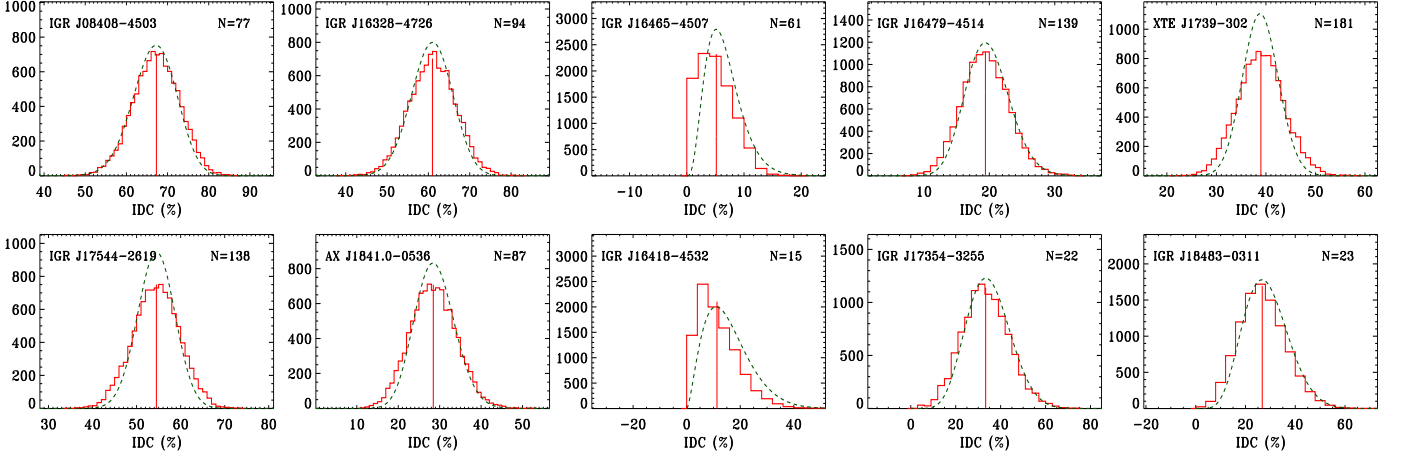


Fig. 1. Distribution of IDC values derived from 10^4 bootstrap simulations (red), each drawn from a sample of size N . The solid vertical line marks the simulated sample mean from Eq. (12). The dashed (green) lines are the curves described by Eq. (7) in the case of a uninformative prior ($a = b = 1$).

in agreement with the asymptotic limit of a normal distribution.

For a class of sources consisting of a small number of individuals ($N_{\text{src}} \lesssim 50$) the prior $p(\mu)$ is unknown, so only an uninformative prior can be used in Eq. (4). Such is the case of SFXTs ($N_{\text{src}} = 10$), which will be detailed in Sect. 3, and for which Eq. (7) can only be used with $a = b = 1$.

To this end, we provide (on-line only) R-language, IDL, and C-language programs that, given N and DC as calculated according to Eq. (2), provides the 68.3 %, 95.4 %, and 99.7 % c.i. for the theoretical distribution (Eq. 7).

On the contrary, when $N_{\text{src}} > 50$ the prior can be obtained from the observed distribution of the DCs of all sources by fitting it with the Beta function in Eq. (6) with free parameters a and b . In this case, Eq. (6) turns out to be particularly useful for the newly discovered sources even with relatively few available observations. Through the prior, one can further constrain the DC of a new source by exploiting the information (the fitted values of a and b) acquired on the DC distribution derived from the sources of the same class previously observed.

3. Evaluating duty cycles with Monte Carlo bootstrap simulations

Once the best available measurement, $DC(N)$, has been obtained from a set of N independent observations, one needs to assess its associated error. The DC determinations obtained by accumulating increasing observing time are not independent; therefore, the dataset cannot be used to directly determine the error on DC. Furthermore, the datasets can be so poor that the hypothesis of normal errors does not apply. The standard approach, also validating *a posteriori* our derivation in Sect. 2, is to perform Monte Carlo simulations.

As a test case, we consider the *Swift*/XRT monitoring campaigns on the ten SFXTs reported in Table 1, discussed in full by Romano et al. (2014) who calculate the IDCs according to Eq. 2. Table 1 (Cols. 1–5) reports the main properties of the sample. The data were divided in *i*) yearly campaigns (Y), a casual sampling of the X-ray light curve of an SFXT at a resolution of $P_{\text{samp}} \sim 3\text{--}4$ d over a $\sim 1\text{--}2$ yr baseline (for these, $P_{\text{samp}} \geq P_{\text{orb}}$); and *ii*) orbital campaigns (O), that sample the light curve intensively with $P_{\text{samp}} \ll P_{\text{orb}}$ so that the phase space is uniformly

observed within one (or a few) P_{orb} . Further details can be found in Romano et al. (2014).

In order to determine the expectation value of IDC and its error, we performed Monte Carlo bootstrap simulations (Efron 1979, 1994). We created $M = 10^4$ simulated data sets, drawn from the observed sample of size N with a simple sampling (with replacement, or uniform probability). We calculated M values of IDCs (simulated sample) according to Eq. (2). The simulated sample mean and standard variance (Table 1, Col. 9) are

$$\overline{IDC}_{\text{sim}} = \frac{1}{M} \sum_{k=1}^M IDC_{\text{sim}}(k), \quad (12)$$

$$s_{\text{sim}}^2 = \frac{1}{M-1} \sum_{k=1}^M (IDC_{\text{sim}}(k) - \overline{IDC}_{\text{sim}})^2. \quad (13)$$

In Fig. 1 we show, superposed on the simulated sample distributions (solid red curves), the simulated sample mean $\overline{IDC}_{\text{sim}}$ (vertical line), and the theoretical expectations (dashed green curves) described by Eq. (7). We find that $s_{\text{sim}} = 2.9\text{--}6\%$ for the yearly campaigns and $s_{\text{sim}} = 8.0\text{--}10.4\%$ for the orbital ones.

The standard c.i., defined by the integral of the probability function (i.e. the simulated distributions), the cumulative probability function,

$$F(x) = \int_{-\infty}^x IDC_{\text{sim}}(x') dx', \quad (14)$$

can be calculated from

$$F(x_{1,1}) = \frac{1-c1}{2}; \quad F(x_{2,1}) = \frac{1+c1}{2}; \quad \text{and } c1 = 0.6827, \quad (15)$$

$$F(x_{1,2}) = \frac{1-c2}{2}; \quad F(x_{2,2}) = \frac{1+c2}{2}; \quad \text{and } c2 = 0.9545, \quad (16)$$

$$F(x_{1,3}) = \frac{1-c3}{2}; \quad F(x_{2,3}) = \frac{1+c3}{2}; \quad \text{and } c3 = 0.9973. \quad (17)$$

3.1. IDC as a function of sample size

We can now determine the expected IDC value for a given observed sample size via additional Monte Carlo bootstrap simulations. For each of the sources monitored with yearly campaigns, we created $M = 10^4$ datasets drawn from the first

$S = 10, 20, 30, \dots, N$ observed points, with a simple sampling (with replacement, or uniform probability). The simulated sample mean \overline{IDC}_S and the standard deviation s_S were calculated similarly to Eq. (12)–(13).

Figure 2 shows $\overline{IDC}_S \pm s_S$ as a function of the sample size S . The last point (filled triangle) is the simulation for N points for which $\overline{IDC}_N = \overline{IDC}_{\text{sim}}$ and $s_N = s_{\text{sim}}$ (Eqs. 12 and 13). The red-orange-yellow bands mark the 68.3 %, 95.4 %, and 99.7 % c.i. for the simulated distribution as derived from Eqs. (15)–(17). We note the excellent correspondence between the 68.3 % c.i. (red band) and the simulated sample standard deviation s_N (the error-bar on the simulation for N points), as expected from a normal distribution. The green bands (from dark to light green) mark the 68.3 %, 95.4 %, and 99.7 % c.i. for the theoretical distribution in Eq. (7) also reported in Table 1, Cols. 6–8.

We define S_a as the minimum S value for which $IDC(S)$ is considered acceptable, that is the number of observations required in order to satisfy both conditions:

$$\overline{IDC}_S \in [\overline{IDC}_{\text{sim}} - s_{\text{sim}}, \overline{IDC}_{\text{sim}} + s_{\text{sim}}]$$

$$\overline{IDC}_S \pm s_S \in [\overline{IDC}_{\text{sim}} - 2 s_{\text{sim}}, \overline{IDC}_{\text{sim}} + 2 s_{\text{sim}}].$$

The values of S_a thus determined are reported in Table 1, Col. 10, and they range between 40 and 80 observations, depending on the source.

Similarly, for each of the sources monitored with orbital campaigns, we created $M = 10^4$ datasets drawn from $S = 5, 10, 15, 20, \dots, 70$ observed points, thus also extrapolating the observed sample to determine how many additional observations are required to significantly lower the uncertainty s_S . We find that for about 70 observations $s_S = 3.6$ – 5.8 %, thus comparable to those found for the yearly monitoring campaigns.

These findings can easily be used for planning future observations.

4. Conclusions

As an alternative and considerably less time-consuming approach than Monte Carlo bootstrap simulations, we derived the theoretical Bayesian expectation value for a duty cycle and its error based on a finite set of independent observational data points. We have applied our findings to the specific case of the inactivity duty cycle of SFXTs, as one of the available examples of two-state sources. For SFXTs we have compared the theoretical expectations with both the observed values and with the IDCs and their errors obtained from Monte Carlo simulations, as an *a posteriori* validation of the Bayesian treatment.

Our treatment, however, is more general than the simple case we considered and can be applied to all independent observations of two-state sources, such as AGNs, X-ray binaries, etc., suitable for a meaningful DC determination. In particular, the strength of this approach becomes evident when considering a well-populated class of sources ($N_{\text{src}} \geq 50$) for which, the parameters a and b can be obtained by fitting the distribution of the observed DCs for all sources in the class with the Beta function in Eq. (6), thus fully characterizing the prior. Then, whenever a new source in the same class is observed for relatively few observations, the knowledge of the prior derived from the whole class can be utilized to further constrain the DC of this still poorly studied individual source by adopting the a and b of the class.

Acknowledgements. We thank A. Stamerra, P. Esposito, V. Mangano, and E. Bozzo for helpful discussions. CG acknowledges PRIN MIUR project on "Gamma Ray Bursts: From Progenitors to Physics of the Prompt Emission Process," PI: F. Frontera (Prot. 2009 ERC3HT). LD thanks Deutsches Zentrum für

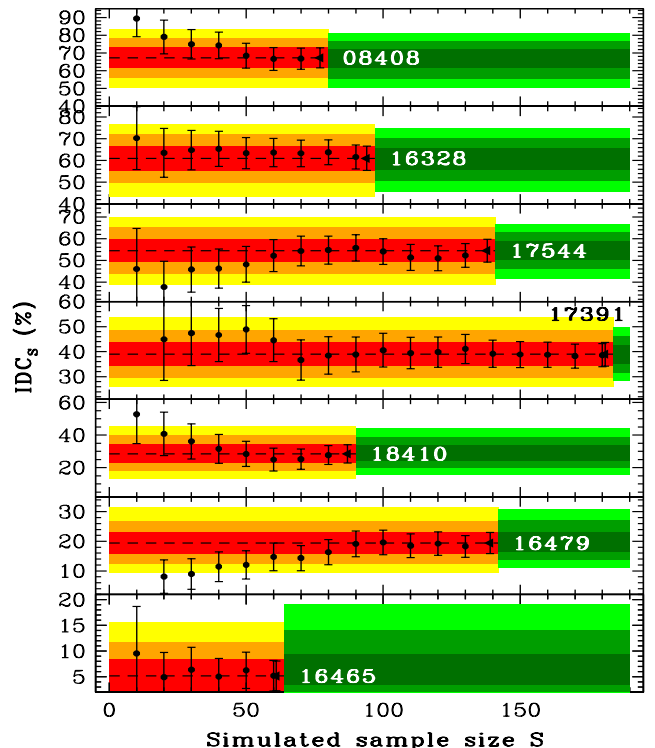


Fig. 2. Simulated sample means and their errors $\overline{IDC}_S \pm s_S$ as a function of sample size S (points) for the yearly sample. The last point (filled triangle) is the simulation for N points. The shaded areas mark the 68.3 % (red), 95.4 % (orange), and 99.7 % (yellow) confidence intervals for IDC_{sim} (see Sect. 3) and from Eq. (7) (dark green, green, light green, respectively; see Table 1 in the case of a uninformative prior, $a = b = 1$).

Luft und Raumfahrt (Grant FKZ 50 OG 1301). We also thank the referee for comments that helped improve the paper. The *Swift*/XRT data were obtained through target of opportunity observations (2007-2012; contracts ASI-INAF I/088/06/0, ASI-INAF I/009/10/0) and through contract ASI-INAF I/004/11/0 (2011-2013, P.I.P. Romano).

References

- Ackermann, M., Ajello, M., Allafort, A., et al. 2011, *ApJ*, 743, 171
- Bishop, C. M. 2006, *Pattern Recognition and Machine Learning*, ed. Jordan, M., Kleinberg, J., & Scholkopf, B.
- Burrows, D. N., Hill, J. E., Nousek, J. A., et al. 2005, *SSR*, 120, 165
- Efron, B. 1979, *Ann. Statist.*, 7, 1
- Efron, B. 1994, *An introduction to the bootstrap* (New York: Chapman & Hall)
- Fragos, T., Kalogera, V., Willems, B., et al. 2009, *ApJL*, 702, L143
- Gehrels, N., Chincarini, G., Giommi, P., et al. 2004, *ApJ*, 611, 1005
- Henry, G. R. & Paik, H.-J. 1969, *Nature*, 224, 1188
- Jorstad, S. G., Marscher, A. P., Mattox, J. R., et al. 2001, *ApJ*, 556, 738
- Kneivitt, G., Wynn, G. A., Vaughan, S., & Watson, M. G. 2014, *MNRAS*, 437, 3087
- R Core Team. 2014, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria
- Romano, P., Ducci, L., Mangano, V., et al. 2014, *A&A*, 568, A55
- Romano, P., Sidoli, L., Cusumano, G., et al. 2009, *MNRAS*, 399, 2021
- Vercellone, S., Soldi, S., Chen, A. W., & Tavani, M. 2004, *MNRAS*, 353, 890